Working Paper 01-17
Statistics and Econometrics Series 11
March 2001

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

SEMIPARAMETRIC MODELS AND P-SPLINES

Currie, I. and Durbán, M.*

Abstract

P-splines were introduced by Eilers and Marx (1996). We consider semiparametric models where the smooth part of the model can be described by P-splines. A mixed model representation is also considered. We set a simple strategy for the choice of P-spline parameters, ndx, bdeg and pord, and discuss the use of various criteria for smoothing parameter selection. We illustrate our remarks with the analysis of a randomised block design.

Keywords: AIC; GCV; mixed models; P-splines; randomised blocks; semiparametric models.

*Currie, Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland, UK.; Durbán, Department of Statistics and Econometrics; Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganés (Madrid); e-mail: mdurban@est-econ.uc3m.es

Semiparametric models and P-splines

Iain Currie[1] and Maria Durbán[2]

[1] Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland
[2] Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Madrid, Spain

Abstract: P-splines were introduced by Eilers & Marx (1996). We conside semiparametric models where the smooth part of the model can be described by P-splines. A mixed model representation is also considered. We set out a simple strategy for the choice of P-spline parameters *ndx*, *bdeg* and *pord*, and discuss the use of various criteria for smoothing parameter selection. We illustrate our remarks with the analysis of a randomised block design.

Keywords: AIC; GCV; mixed models; P-splines; randomised blocks; semiparametric models.

1 Introduction

The introduction of B-splines with penalties, known as P-splines, by Eilers and Marx (1996) provided another approach to non-parametric modelling. P-splines have many attractive properties amongst which we mention only their flexibility, their ease of computation and their connection to smoothing splines and polynomial regression. The examples discussed by Eilers and Marx were such that the fitted smooth function was the principal focus of the analysis. However, in semiparametric models it is common for the underlying smooth function to be a nuisance parameter; it is the effects of the regressor variables that are of interest.

In this paper we use P-splines to model smooth background variation in a semiparametric model. The user of P-splines has many choices to make: the domain, the degree of the P-spline, the order of the penalty, and the number and location of the knots. Once a particular P-spline has been chosen there is a second choice, the method of smoothing parameter selection. Eilers and Marx recommend using AIC or GCV but there is considerable evidence that their application in the semiparametric setting often leads to undersmoothing; see for example, Hurvich, et al. (1998). We will consider a number of alternatives to AIC and GCV that might be more suitable in the semiparametric context.

As an example we consider a randomised block experiment with a large number of varieties and low replication. The classical model assumes that block effects are fixed. However, there is a good argument for taking block effects as random, and, in the same fashion, we consider smoothly varying block effects as random and express the semiparametric model as a mixed model. With this formulation we can choose the level of smoothing by residual maximum likelihood (REML).

2 Semiparametric models and *P*-splines

Suppose the variable y depends smoothly on a single variable x then the nonparametric model for y can be written $y = f(x) + \epsilon$ where $f(\cdot)$ is a smoothly varying function and ϵ is an error term with variance σ^2. Eilers and Marx (1996) make two assumptions: first, with data (x_i, y_i), $i = 1, \ldots, n$, they assume that $y_i \approx \sum a_j B_j(x_i)$ where $B_j(\cdot)$ is a set of B-splines; second, they suppose that the coefficients of adjacent B-splines satisfy certain smoothness conditions which can be expressed in terms of finite differences of the a_i's. Thus, from a least squares perspective, the coefficients, a, are chosen to minimise

$$S(a) = (y - Ba)'(y - Ba) + \lambda a'D'Da \tag{1}$$

where D is a difference matrix and λ is a penalty. For given λ, the solution to this optimisation problem satisfies

$$(B'B + \lambda D'D)\hat{a} = B'y \tag{2}$$

and then $\hat{y} = B\hat{a} = Hy$ where H is the hat-matrix:

$$H = B(B'B + \lambda D'D)^{-1}B'. \tag{3}$$

The extension to the semiparametric case is straightforward. Suppose we have the model $y = X\beta + f(x) + \epsilon$ then we choose β and a by minimising

$$S(a, \beta) = (y - X\beta - Ba)'(y - X\beta - Ba) + \lambda a'D'Da. \tag{4}$$

The solution to this optimisation problem satisfies

$$\begin{bmatrix} X'X & X'B \\ B'X & B'B + \lambda D'D \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X' \\ B' \end{bmatrix} y \tag{5}$$

with hat-matrix H_X given by

$$H_X = H + (I - H)X[X'(I - H)X]^{-1}X'(I - H) \tag{6}$$

where H is defined in (3).

In the case of a randomised block design with r blocks we want to fit a separate smooth function Ba_i, $i = 1, ..., r$, to model underlying fertility in each block. We can either use a distinct penalty λ_i for each block in which case we minimise

$$S(a, \beta) = \sum_{i=1}^{r} (y - X_i\beta - Ba_i)'(y - X_i\beta - Ba_i) + \sum_{i=1}^{r} \lambda_i a_i' D' D a_i, \qquad (7)$$

or a common penalty λ across blocks, in which case we minimise (7) with $\lambda_i = \lambda$. In both cases we obtain a solution of the form (5) but with B and D replaced by block diagonal matrices, the number of such blocks being equal to r, the number of blocks in the design.
One practical point is that, since $B1 = 1$, it follows that $H1 = 1$, and so 1 must not be in the span of X; in particular, intercept and block effects must not be fitted in X.

3 *P*-splines as mixed models

In order to express the nonparametric trend Ba as the sum of a fixed and random effect we need to write $a = Wb + Zu$ where $[W : Z]$ is square and non-singular and W and Z are such that when $Wb+Zu$ is substituted for a in (2) a set of mixed model equations result. A suitable choice of W and Z is as follows: let $w' = (1, 2, \ldots, k)$ where k is the number of columns of B and define $W = (1, w, w^2, \ldots, w^{q-1})$ where q is the order of the penalty; define $Z = D'(DD')^{-1}$. Then $[W : Z]$ is square and non-singular and furthermore equation (2) reduces to the mixed model equations for the mixed model

$$y = BWb + BZu + \epsilon \qquad (8)$$

where $u \sim \mathcal{N}(0, \sigma_u^2 I)$ is a random effect and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The same transformation reduces (5) to the mixed model equations for

$$y = X\beta + BWb + BZu + \epsilon \qquad (9)$$

where β and b are fixed effects and u is a random effect. The smoothing parameter $\lambda = \sigma^2/\sigma_u^2$ and so λ can be chosen by maximising the residual loglikelihood $\ell(\sigma^2, \lambda)$

$$-\tfrac{1}{2}\log|\Sigma| - \tfrac{1}{2}\log|X'\Sigma^{-1}X| - y'(\Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1})y \quad (10)$$

where, from (9), X is $[X : BW]$ and $\Sigma = \sigma^2(I + \lambda^{-1}BZZ'B')$.

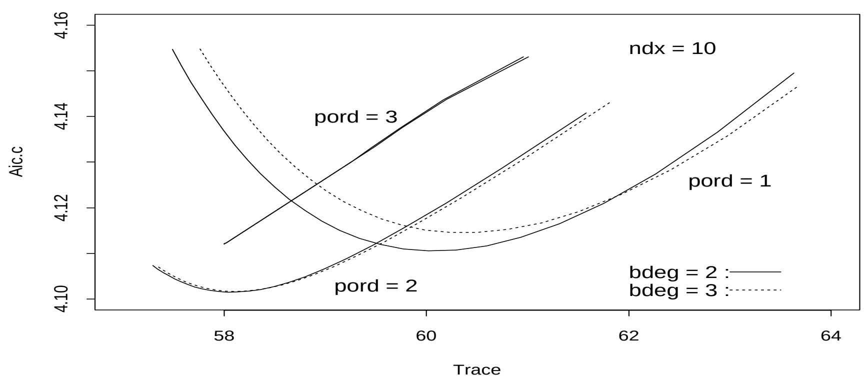

FIGURE 1. AIC_C against $\mathrm{Trace}(H_X)$ for $ndx = 10$, $bdeg = 2$, 3, $pord = 1$, 2, 3

4 An example

4.1 Choice of *ndx*, *bdeg* and *pord*

We use data on wheat yields from a trial conducted in Mexico with a randomised complete block design; see Besag and Higden (1999). A plot of the residuals from the standard variety and block model against plot number provides clear evidence of a fertility trend in each block. We examine the effectiveness of a number of parameterisations of P-splines at controlling trend in these data. In the notation of Hurvich, et al. (1998) we choose λ to minimise functions of the form $\log(\hat{\sigma}^2) + \psi(H_X)$ where $\hat{\sigma}^2 = y'(I - H_X)^2 y/n$, $\psi(H_X)$ is a penalty function and H_X is the hat-matrix defined in (6). Let $t = \mathrm{tr}(H_X)/n$. We consider the following choices of $\psi(H_X)$: $-2\log(1-t)$ (GCV), $2t$ (AIC), $-\log(1-2t)$ (Rice), $(2\mathrm{tr}(H_X)+2)/(n-\mathrm{tr}(H_X)-2)$ (AIC_C), and $2/(1-t)$ AIC_L. Note that AIC_L is the limit of AIC_C as $n \to \infty$ with $\mathrm{tr}(H_X)/n$ finite.
Following the notation of Eilers and Marx (1996) we denote the number of intervals per block by ndx, the degree of the P-spline by $bdeg$ and the order of the penalty by $pord$. How does one decide on suitable values for ndx, $bdeg$ and $pord$? In the simple smoothing situation Eilers and Marx recommend plotting AIC against the trace of H. However, this is unlikely to perform well in the semiparametric setting, so instead, we plot the AIC_C criterion of Hurvich, et al. (1998) against $\mathrm{tr}(H_X)$. Figure 1 suggests that, with $ndx = 10$, we should take $pord = 2$ and $bdeg = 2$ or 3. For simplicity we will fix on $bdeg = 3$. Now, with $pord = 2$ and $bdeg = 3$, Fig. 2 gives the same plot for $ndx = 10$, 25 and 50. The basic philosophy of P-spline methodology is that the penalty will look after the number of knots so we would not expect much difference between models with different ndx;

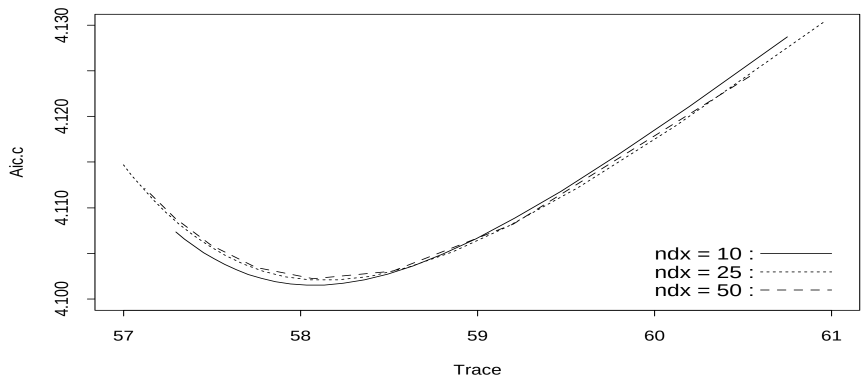

FIGURE 2. AIC_C against Trace(H_X) for $ndx = 10$, 25, 50, $bdeg = 3$, $pord = 2$

this is borne out by Fig. 2. We will fix on $ndx = 10$ for two reasons: first, there is a small preference for $ndx = 10$ in Fig. 2; second, there is a substantial computational advantage with $ndx = 10$ over both of the other values. For example, with $ndx = 50$, the unrestrained model $y \approx X\beta + Ba$ has 208 parameters compared to 88 with $ndx = 10$; of course, with the appropriate penalty both these models have an effective dimension of approximately 58, as can be seen in Figs. 1 and 2.

4.2 Smoothing parameter selection

We now consider smoothing parameter selection within the family of models with $ndx = 10$, $bdeg = 3$ and $pord = 2$. Figure 3 shows the fitted trend chosen by AIC_C ($\lambda = 29$, $\mathrm{Tr}(H_X) = 58$). Other criteria can result in very different fits: AIC ($\lambda = 0.005$, $\mathrm{Tr}(H_X) = 82$) is close to interpolation; GCV ($\lambda = 1.9$, $\mathrm{Tr}(H_X) = 63$); the mixed model (REML) gives $\lambda = 9.3$, $\mathrm{Tr}(H_X) = 60$. Figure 4 is a standardised residual plot after fitting varieties and trend again with $\lambda = 29$. The plot looks satisfactory.

This analysis assumes a common λ for each block but it is possible that the trend in each block might require smoothing on a different scale, i.e. distinct λ. If we adopt the mixed model formulation we can test $H_0 : \lambda_i = \lambda$ v $H_1 : \lambda_i \neq \lambda$ with REML. Under H_0, we have $\lambda = 9.3$, while $\lambda_1 = 11.8$, $\lambda_2 = 14.8$ and $\lambda_3 = 4.4$ under H_1; the residual loglikelihood increases by 0.2 so there is little evidence that distinct λ are required.

We look now at the effect on variety means of modelling trend. Variety 18 is the variety which is most advantaged by the randomisation of the trial, in the sense that it has the largest downward adjustment (irrespective of the choice of λ). Figure 5 shows the adjustments for $0 < \lambda < 60$. Also shown are the adjustments for variety 48, the most disadvantaged variety. The adjustments to these varieties are very substantial and have a marked effect on their rankings. We note also that the size of the adjustments depends very much on the choice of λ with AIC and GCV resulting in much larger adjustments than the other criteria.

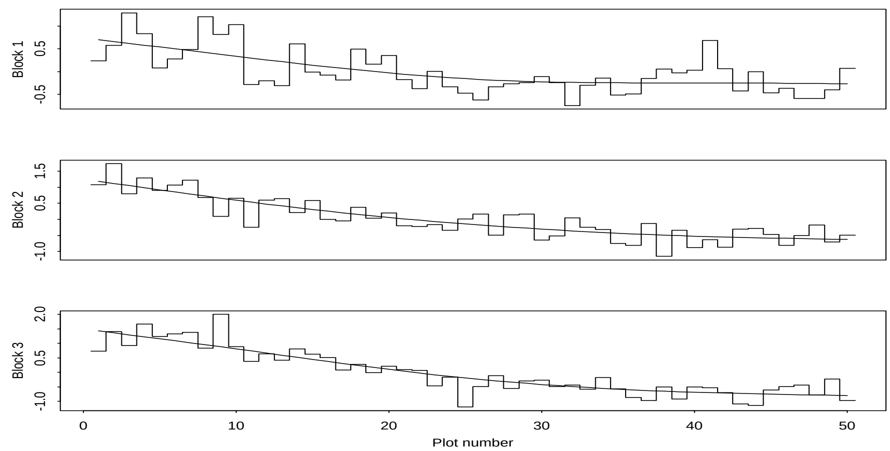

FIGURE 3. Fitted trend and residuals $y - X\hat{\beta}$ fitted by AIC_C ($\lambda = 29$) with $ndx = 10$, $bdeg = 3$ and $pord = 2$

Figure 6 shows how the average standard error of differences (SED) of variety means depends on λ. We note the following: the SED for the simple block model is 0.59 so failure to remove the strong trend results in serious overestimation of the SED; the average SED does not vary greatly with λ over the displayed range; the average SED given by AIC_C (and AIC_L) and REML are very close; as $\lambda \to \infty$ the average SED tends to 0.39, the value given by fitting a linear effect for blocks.

5 Concluding remarks

In this short paper we have considered semiparametric models where the smooth part of the model can be described by P-splines. We have set out a

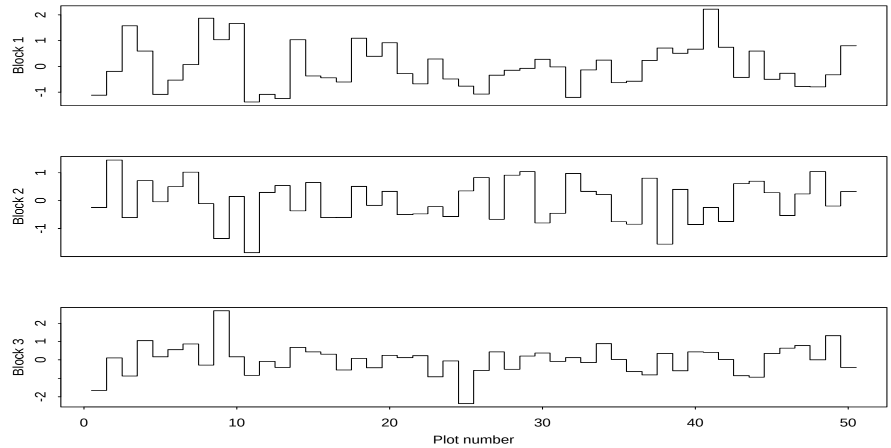

FIGURE 4. Residuals $(y - X\hat{\beta} - \hat{f})/s$ fitted by AIC_C ($\lambda = 29$) with $ndx = 10$, $bdeg = 3$, $pord = 2$ and $s^2 = y'(I - H_X)^2 y/(n - \mathrm{tr}(H_X))$

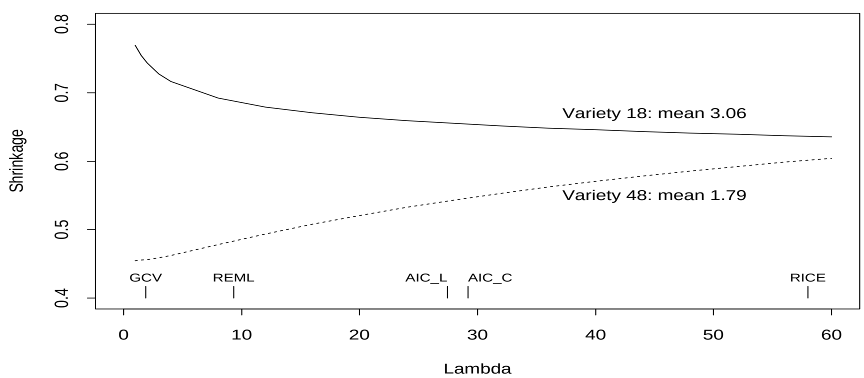

FIGURE 5. Shrinkage of variety means for advantaged variety 18 and disadvantaged variety 48; $ndx = 10$, $bdeg = 3$, $pord = 2$.

simple strategy for the choice of P-spline parameters ndx, $bdeg$ and $pord$, and discussed the use of various criteria for smoothing parameter selection. In our example, we described the effect of the smoothing parameter on both the estimates of variety means and on their average standard errors. One feature of P-splines which we find particularly appealling is their ease of calculation. Equation (5) is at the heart of all calculations and, for given λ, this is a simple linear system.

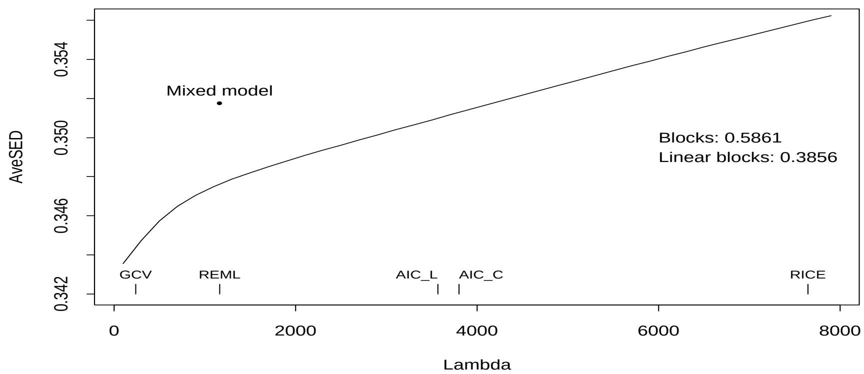

FIGURE 6. Average standard error of variety differences with $ndx = 50$, $bdeg = 3$, $pord = 2$

References

Besag, J. & Higden, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **61**, 691-746.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11**, 89-121.

Hurvich, C. M., Simonoff, J. S. & Tsai, C-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy. Statist. Soc. Ser. B* **60**, 271-293.